基于深度玻尔兹曼机的乐器分类问题研究

周　畅，米红娟

(兰州财经大学 信息工程学院，兰州 730020)

摘　要：应用传统浅层模型处理乐器分类任务存在非线性拟合能力较差的问题，使分类准确率得不到有效保证，有必要引入深度学习方法提升复杂任务的非线性建模能力。将深度玻尔兹曼机作为特征提取器提取表达能力更强的数据特征，分别以 SVM 与 Softmax 分类器作为深度神经网络的顶层设置形成 DBM+SVM 组合模型与 DBM+Softmax 组合模型，引入平均场理论和动量项因子优化网络训练过程。将上述两组模型及单一 SVM 分类器在 5 类乐器音频数据上进行对比实验，两种深度学习组合模型的分类准确率分别达到 89.29%和 87.5%，与传统浅层分类方法 SVM 的 73.21%的准确率相比优势明显。实验结果表明深度玻尔兹曼机在乐器分类领域的应用颇具前景。

关键词：深度玻尔兹曼机；乐器分类；深度学习；平均场理论；动量项

中图分类号：TP391.4　　doi: 10.3969/j.issn.1001-3695.2018.01.0022

Research on musical instrument classification based on deep Boltzmann machine

Zhou Chang, Mi Hongjuan

(*Department of Information Engineering*, *Lanzhou University of Finance & Economics*, *Lanzhou* 730020, *China*)

Abstract: The application of traditional shallow model to instrument classification task had the problem of poor nonlinear fitting ability, so that the accuracy of classification was not guaranteed effectively. It was necessary to introduce deep learning method to improve the nonlinear modeling ability of complex tasks. Deep Boltzmann machine was used as feature extractor to **Abstract:** more expressive deep learning features. SVM and Softmax classifier were respectively used as top layer of deep neural network to form DBM + SVM and DBM + Softmax combined model. Besides, the mean field theory and momentum factor were introduced to optimize the network training process. The above two sets of models and single SVM classifier was compared on 5 kinds of musical instruments audio data. The classification accuracy of the two types of deep learning combination models reached 89.29% and 87.5%respectively, compared with the accuracy of the traditional shallow classification method SVM of 73.21% . The experimental results show that the application of deep Boltzmann machine in the field of musical instrument classification is very promising.

Key words: deep Boltzmann machine; instrument classification; deep learning; mean field theory; momentum term

0　引言

乐器分类任务就是要求分类系统能够根据输入的音乐信息判断演奏该乐曲的乐器。这一任务对于音乐信息检索来说至关重要，因为目前网络上有大量缺乏数据标签的音频文件，尤其是人工标签中的“乐器”属性值经常处于缺省状态，这就意味着用户在以乐器名称作为关键字搜索时，基于文本的音乐检索很难返回准确的结果信息。文献[1]提取了乐器音频数据的 MFCC 系数以及 LPC 系数(线性预测系数)作为特征量，并结合 GMM 构建声学模型，采用 SVM 作为分类器训练得到 9 种乐器的分类准确率仅为 70%，这就是乐器分类较为早期的研究。近年来，Petros[2]从乐器音频数据中提取调幅-调频调制特征量进行分析并应用于乐器识别问题，取得了超过 70%的准确率。文献[3]通过提取和分析由 10 种西方乐器演奏的大量音乐片段的频谱特征和倒谱特征，并通过实验证明用倒谱特征训练的 SVM 分类器的分类准确率较频谱特征的训练结果有较大幅度的提升。特征量的表达能力对训练分类器来说至关重要，构建高性能的混合特征量也逐渐成为一种趋势[4]。文献[5]在每一个组建 DBN 的 RBM 的可视层和隐含层间设置一个用以防止过拟合的 Dropout 层来提升模型的泛化能力，并引入一个能够表征权值影响大小的动量用以平衡模型在训练速度和稳定性上的关系，将其应用于中国传统乐器的分类识别问题，取得了较高的分类准确率。而 Dropout 方法在神经网络中的早期应用是由 Hinton 等人[6]提出的，通过在训练过程中抑制半数特征检测器的工作

收稿日期：2018-01-13；**修回日期：**2018-03-20

作者简介：周畅（1992-），女，新疆乌鲁木齐人，硕士研究生，主要研究方向为机器学习、深度学习、(ZhouCh28@foxmail.com)；米红娟（1964-），女，甘肃宁县人，教授，学士，主要研究方向为数据挖掘、机器学习、智能决策.

来达到缓解模型过拟合的效果。

深度学习方法可以有效训练大量无标签数据，并从中提取表达能力较强的特征量。因此，基于深度玻尔兹曼机的乐器分类方法是一种基于内容的分类模式，应用深度学习技术的音频分类系统是更贴近用户实际需求且能够有效降低数据管理成本的分类系统。

1　相关理论

1.1　深度学习的概念

深度学习的概念起源于人工智能的相关研究，从简单的单层感知机发展到隐层数目不受限的多层感知机，其概念已逐步形成。G.E. Hinton[7]等在 2006 年发表于《科学》杂志上的一篇文章中采用逐层预训练的方式有效解决了传统反向传播算法易陷入局部最优的问题，由此开启了深度学习的研究热潮。文中成功构建了具有多个隐含层的深度信念网络(DBN)，其研究思路包含两个阶段：第一阶段的无监督学习。用无标签初始数据依次对神经网络的每一层进行训练，前一层的训练结果作为后一层的输入。在训练整个深度网络存在困难的情况下，采用将多层神经网络拆分成为独立的多个单层结构进行训练的思路非常有效。第二阶段的有监督学习。采用反向传播算法，利用有标签数据自上而下地对整个网络的权值进行微调。

在网络结构的构建中，如果用 m 层结构能够简易地表达一组信息，那么当采用 m-1 层结构的时候，则可能需要指数级数量的参数才能够达到同样的效果，同时模型的泛化能力也会大打折扣[8,9]，而深度学习突破了网络层数对训练有效性的限制，找到并不断优化处理复杂网络问题的方法[10]。另外，与适用于浅层结构的学习算法相比，深度学习方法主要依靠计算机自主完成学习过程，提取表达能力更强的特征并用于最终的分类或预测，以期达到更好的效果[11]。

1.2　深度玻尔兹曼机

深度玻尔兹曼机(DBM)是玻尔兹曼机(BM)的一种特殊形式，与构建深度信念网络的受限玻尔兹曼机(RBM)的相似之处在于只有相邻两个层次间的神经元可以连接，同一层次或不相邻层次中的神经元之间没有连线；而二者的差异体现在 RBM 只包含一个隐含层，而 DBM 可以有多个隐含层。DBM 虽然在层级结构上类似 DBN，都是多层次的生成式模型，但 DBM 是无向图模型[12]，其网络结构中的所有连接都是无向的，而在 DBN 网络中只有最高两层的连接是无向的，二者的差异如图 1 所示。

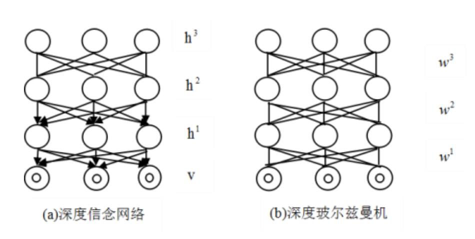


图 1　DBN 与 DBM 的网络结构图

假设一个深度玻尔兹曼机含有 m 个隐含层，令其隐含层向量为$h_k=(h_{k,1},h_{k,2},\cdots,h_{k,n_k})^T$，k=1，2，…，m。可视层向量 v 和第一个隐含层间的连接权值向量表示为W^1，第 k-1 个隐含层与第 k 个隐含层之间的权值向量表示为W^k，此时，2≤k≤m。令b_v表示可视层的偏置，b_h^k表示第 k 个隐含层的偏置，1≤k≤m。可以令h_0=v，$b_h^0=b_v$，则可将该 DBM 的能量函数定义为

$$\begin{aligned}\mu(v,h_1,h_2,\cdots,h_m|\theta)&=-v^T(W^1)^Th_1-(h_1)^T(W^2)^Th_2-\cdots\\&(h_{m-1})^T(W^m)^Th_m-v^Tb_v-(h_1)^Tb_h^1-\cdots-(h_m)^Tb^m\\&=-\sum_{k=1}^m(h_{k-1})^T(W^k)^Th_k-\sum_{k=0}^m(h_m)^Tb^m\end{aligned}\quad(1)$$

其中:$\theta=\{W^k,b_v,b_h^k\}\ (1\le k\le m)$ 是模型参数。

由已知能量函数的定义，可进一步推导 DBM 关于可视层向量 v 的概率分布如下:

$$p(v|\theta)=\frac{1}{Z(\theta)}\sum_{h_1,h_2,\cdots,h_m}exp\left(-\mu(v,h_1,h_2,\cdots,h_m|\theta)\right)\quad(2)$$

其中:Z 为归一化因子（在此使用剖分函数）。

结合概率图模型的结构及相关理论，还可对 DBM 推导出如下条件概率计算公式：

$$\begin{cases}p(v|h_1,\theta)=\prod_j p(v_j|h_1,\theta)\\p(v_j=1|h_1)=sigmoid(\sum_i w_{ij}^1h_{1i}+b_{v,j})\end{cases}\quad(3)$$

$$\begin{cases}p(h_k|h_{k-1},h_{k+1})=\prod_i p(h_{ki}|h_{k-1},h_{k+1})\\p(h_{ki}=1|h_{k-1},h_{k+1})=sigmoid(\sum_j w_{ij}^kh_{k-1,j}+\sum_j w_{ij}^{k+1}h_{k+1,j}+b_i^k),1\le k\le m-1\end{cases}\quad(4)$$

$$\begin{cases}p(h_m|h_{m-1})=\prod_i p(h_{mi}|h_{m-1})\\p(h_{mi}=1|h_{m-1})=sigmoid(\sum_j w_{ij}^mh_{m-1,j}+b_i^m)\end{cases}\quad(5)$$

$$\mathrm{sigmoid}(x)=\frac{1}{1+exp(-x)}\quad(6)$$

其中：sigmoid(x)是激活函数，一方面规范神经元的输出结果，另一方面是为神经网络的处理加入非线性因子，毕竟实际应用中处理的绝大部分是线性不可分问题，作用于神经元的激活函数恰能有效处理类似情况。

近年来，深度玻尔兹曼机在模式识别、自然语言处理、图像及语音识别等领域逐渐表现出与其他算法相比的竞争优势。蒋文等人[13]将 DBM 理论用于改进典型相关分析算法(CCA),通过 DBM 方法提取可视层特征与隐含层特征，并在此基础上结合有标签数据构建串行融合特征集与并行特征融合集，最终在多个数据库的验证中展现了优于其他几种传统算法的识别性能。文献[14]将输出结构的高阶相关性纳入考虑范围，利用深度玻尔兹曼机为图像分割任务提供先验形状信息，提出了用以捕获输出空间特征表示的条件深度玻尔兹曼机模型(CDBM)，并在 Part Labels 人脸数据集上进行有效性验证，取得了优于传统算法的分割效果。

2　模型搭建

2.1　平均场理论

平均场理论最初产生于统计力学领域，但目前已被广泛应

用于网络传播机制[15]、无线自组织网络[16]和站点优化[17]等其他领域。平均场理论的基本思路是将多体问题转换为单体问题，用单体有效场来替代其他所有单个节点对目前研究节点的个性化影响，因此该理论指导研究人员在处理问题时不依赖于特殊点的作用，而是用整个环境的综合作用近似替代单个影响粒子的效果。

在本文的预处理标准下，音频数据经过分帧加窗操作后，10 s 时长的文件可以被切割成 999 条数据记录，这些表征内涵的音乐数据在短时内可以保持能量的相对平稳，但当时间跨度较大时，将会存在明显的能量的差异性分布。因此，本研究在进行深度玻尔兹曼机建模的过程中，将平均场理论引入模型的训练，目的是将一个音乐片段中偏离整体能量均衡的样本点的影响效果削弱，从而使训练得到的网络参数更有利于特征提取。

平均场理论的核心思想是通过随机变量均值的函数近似估计随机变量的函数的均值。将其应用于深度玻尔兹曼机建模的关键点就在于要引入一个具有因子分解结构的近似概率分布 $q(h_k|v)=\prod_i q(h_{ki}=1|v)$,通过计算$q(h_{ki}=1|v)$来逼近后验概率$p(h_{ki}=1|v)$，并结合平均场理论估计

$h_{ki}\approx p(h_{ki}=1|v)\approx q(h_{ki}=1|v)$。估计后验概率的平均场算法如下：

估计后验概率的平均场算法

输入：可视向量 v，权值矩阵W^k、隐含层偏置b_h^k(2≤k≤m)、网络结构信息、激活函数 Sigmoid。

输出：$q(h_{ki}=1|v)$，2≤k≤m，1⩽i⩽n_k。

$h_{1i}^0=sigmoid\left(\sum_j w_{ij}^1 v_j+\sum_j w_{ij}^1 v_j+b_i^1\right)$

$h_{ki}^0=sigmoid\left(\sum_j 2w_{ij}^k h_{k-1,j}^0+b_i^k\right)$ 2≤k≤r-1;

$h_{mi}^0=sigmoid\left(\sum_j w_{ij}^m h_{m-1,j}^0+\sum_j w_{ji}^{m+1} y_j+b_i^m\right)$

for t=1:n **do**

$h_{1i}^t=sigmoid\left(\sum_j w_{ij}^1 v_j+\sum_j w_{ji}^2 h_{2,j}^{t-1}+b_i^1\right)$

$h_{ki}^t=sigmoid\left(\sum_j w_{ij}^k h_{k-1,j}^t+\sum_j w_{ji}^{k+1} h_{k+1,j}^{t-1}+b_i^k\right)$, 2≤k≤m-1;

$h_{mi}^t=sigmoid\left(\sum_j w_{ij}^m h_{m-1,j}^t+\sum_j w_{ji}^{m+1} y_j+b_i^m\right)$

$\varepsilon_k=\sum_i\left(h_{ki}^t-h_{ki}^{t-1}\right)^2$, 1≤k≤m;

If($\partial_1<\partial$ and $\partial_2<\partial$ and $\partial_3<\partial$ and⋯and $\partial_m<\partial$) break;

end for

$q(h_{ki}=1|v)$=h_{ki}^n; 1≤k≤m

2.2　动量项

动量项系数表示在网络训练过程中，本次训练引起的权值变化量受前一次训练引起的权值变化量的影响程度。动量项系数越大，意味着历史训练对当前训练产生的影响就越大，但这种影响作用是否对整体训练有益则要视具体情况而定。根据动量项的定义，该系数一般被添加在权值修正公式中，如式(7)所示。

$$\Delta w(n+1)=-\rho E'\left(w(n)\right)+\tau(1-\cos\theta)\Delta w(n) \tag{7}$$

其中：Δw代表权值修正量；ρ代表权值学习率；E 代表网络结束一次样本训练时产生的累计误差（常采用均方根误差）；τ代表权值更新过程中的动量项系数；$-E'\left(w(n)\right)=-\frac{\partial E}{\partial w(n)}$, θ代表前一次训练的权值调整量$w(n)$与当前训练的负梯度下降方向之间的夹角[18]。根据式(7)的表达：

a)当$\theta\rightarrow 0$时，前一次训练的权值调整量与当前负梯度方向几乎相同，表明前一次训练的结果对此次训练有较大的正向影响，所以需要加大动量项系数τ的取值来增强这种效果。

b)当$\theta\rightarrow 180°$时，前一次训练的权值调整量与当前负梯度方向相反，表明前一次训练的结果对此次训练有较大的负向影响，因此有必要减小动量项系数τ的取值来削弱这种效果。

在权值修正公式中添加动量项系数能够充分考虑到网络在过去的训练中所累积的学习经验，有效避免网络学习过程中的大规模震荡现象，一定程度上提升网络的训练速度。

2.3　深度玻尔兹曼机的训练

对深度玻尔兹曼机的训练包含两个阶段：

1）逐层贪婪预训练

逐层预训练方法是神经网络突破隐层数目限制，针对传统训练算法易陷入局部最优的有效解决方案。其核心思路是化整为零，将整个网络拆分为多个独立的网络层，自底向上，每次只训练网络的一个层次，前一层的输出作为后一层的输入，逐层依次提取数据特征。尽管针对各个层次的训练是独立进行的，但提取的特征量是连贯的，层次越高，抽取的特征量的概括能力就越强。故而，在逐层贪婪预训练中，特征的抽象程度是逐层递进的。

2）参数微调

第一阶段的逐层预训练是自底向上的一个正向的训练过程，而参数微调阶段则采用神经网络的经典算法——BP 算法，反向调整整个网络的权值系数与各层次的偏置量。当所有样本数据完成一次训练后，首先计算实际输出与期望输出间的差值，该误差值即为网络反向调整的输入数据，整个网络修正就是由训练误差触发的，而调整的方向是使误差减小的方向。

3　仿真实验

本文实验的硬件环境如表 1 所示。

表 1　实验设备参数

名称	参数
处理器	Intel(R)Core™ i7-7700K CPU @ 4.20 GHz 4.20 GHz
内存(RAM)	16 GB
显卡	NVIDIA GeForce GTX 1050 Ti

3.1　数据预处理

目前，在音乐文化领域形成的较为完善的乐器分类体系主要包括三类：中国的古典乐器分类体系、印度《戏艺手册》中描述的乐器分类体系和源自西欧布鲁塞尔音乐学院的乐器分类体系。本研究采用的是应用最广泛的四类划分的欧洲乐器分类

体系。在该乐器分类系统中，可将乐器分为乐体类、革膜类、弦乐类和气柱类，分类依据是导致震动发音效果不同的乐器材质。对上述四类乐器划分情况的描述及具体乐器示例在表 2 中列示。

表 2　欧洲乐器分类体系

乐器类别	类别描述	乐器示例
乐体类	乐器部件的制造材料为无任何张力的材质	锣；钟；木鱼；响板
革膜类	乐器部件的制造材料为有张力的革或膜	鼓类乐器
弦乐类	乐器的震动发音来自于有张力的琴弦	钢琴；提琴；扬琴；竖琴；吉他
气柱类	乐器的震动发音来自于空气柱	长笛；小号；黑管；萨克斯风

本文实验所选乐器为弦乐类乐器中的钢琴、扬琴、小提琴和吉他；以及气柱类乐器中的萨克斯风。实验数据来自于分别由 5 类单独乐器演奏的纯音乐无损文件，应用音乐编辑软件 Cool Edit 对音频文件进行采样，采样率取 44.1khz（即每秒采样 44100 次，该采样率为理论上的 CD 音质界限），将文件切割为 10s 长度的音乐片段，该片段即为预处理环节的输入数据。各类乐器对应的训练数据及测试数据分布情况如表 3 所示。

表 3　实验数据

乐器种类	训练数据（长度为 10s 的音乐片段数）	测试数据（长度为 10s 的音乐片段数）
钢琴	69	11
扬琴	57	9
吉他	60	10
小提琴	63	17
萨克斯风	48	9

其中，具体的数据预处理环节包括预加重、分帧、加窗及快速傅里叶变换。预加重是要加强音频信号的高频部分，即增强信号的高频分辨率。分帧的目的是把较长的数据流切割成较小的片段进行处理；例如本文在 44.1 kHz 的采样率下，一个 10 s 的音乐片段就包含 441 000 个样本点，计算机很难进行一次性整体处理，而另一个重要的原因则是音频信号是一种随时间变化的信号，但在较短的时间间隔内可以认为信号几乎不变（即语音信号的短时平稳性），所以有必要对数据进行分段处理。本文采用的帧长为 20 ms,帧移（前后两帧重叠的部分）为 10 ms，完成分帧处理后，一个 10 s 的音乐片段被划分为 999 个样本数据。加窗是为了在分帧处理后减少由于不满足周期截断而造成的谱泄露误差，从而使时域信号在进行傅里叶变换时尽可能满足周期性要求。本文采用汉明窗函数进行加窗操作，其公式如下：

$$\omega(n)=\begin{cases}0.54-0.46cos\left[\frac{2\pi n}{N-1}\right], & 0\le n\le N-1\\ 0, & \text{其他}\end{cases} \tag{8}$$

傅里叶变换的作用是将语音信号从时域变换到频域，该操作在分帧加窗完成后进行。

3.2　搭建深度玻尔兹曼机模型

本文构建的深度玻尔兹曼机的网络结构为 300-500-1000-500-5，其中，输入层的神经元个数是 300。实际上，单位长度为 10s 的音频文件经过分帧加窗等预处理操作后被分割为了 999×882 的频域数据矩阵，经过多次调整输入神经元数目，发现从一个样本的 882 个幅值中随机选取 300 个值作为输入数据对最终分类结果的影响并不大，但由于神经元数目减少了，网络整体训练速度得到明显提升。文章探讨的是 5 种乐器的分类情况，因此网络输出层的神经元个数为 5；三个隐含层的神经元个数分别为 500,1000 和 500。本文在实验阶段分别尝试了隐含层数目为 2,3,4,的情况，发现仅有 2 个隐含层时，处理该问题需要更多的神经元数目，训练速度及精度均不够理想；而含有 4 个隐含层的神经网络的训练效果与含 3 个隐含层时不相上下，但训练时间明显加长，故最终选择隐含层数目为 3 的深度网络结构展开对比实验。深度玻尔兹曼机模型训练过程中的相关参数如表 4 所示。

表 4　逐层预训练过程网络结构参数

网络层	神经元数目	权值学习率	偏置学习率	初始动量项	最终动量项	迭代次数
RBM1	300-500	0.005	0.05	0.6	0.9	200
RBM2	500-1000	0.001	0.05	0.6	0.9	100
RBM3	1000-500	0.001	0.05	0.6	0.9	100

3.3　实验结果

由于神经网络的权值系数是随机产生的，为排除一次性极端情况的影响，本文在固定一组网络参数的情况下进行 20 次实验并取 20 次输出的平均值作为最终结果。本章设计了三组模型的对比实验，分别是：a)深度玻尔兹曼机作为特征提取器搭配顶层 Softmax 分类器(DBM+Softmax 组合模型)；b)单一分类器支持向量机(对于传统分类算法 SVM 的训练同样是取 20 次实验的平均值结果)；c)深度玻尔兹曼机搭配顶层支持向量机分类器(DBM+SVM 组合模型)。采用上述三种模型分别对表 3 中的 5 类乐器数据进行训练和测试。其中，训练集共有长度为 10s 的音乐片段 297 个，测试集共有同类标准的音乐片段 56 个，三组实验的分类结果如表 5 所示。

表 5　3 组模型的分类结果对比情况

模型	训练集错分音乐片段数	训练集分类准确率	测试集错分音乐片段数	测试集分类准确率
SVM	67	77.44%	15	73.21%
DBM+Softmax	32	89.23%	7	87.5%
DBM+SVM	29	90.24%	6	89.29%

由表 5 中三组模型的实验对比结果可知，DBM+SVM 组合模型对五种乐器的分类准确率最高，为 89.29%；DBM+Softmax 组合模型的分类准确率紧随其后，为 87.5%，前两者的差距在 2%以内；而单一分类器 SVM 的分类准确率则明显低于前两组深度学习组合模型。但将 SVM 作为深度玻尔兹曼机模型的顶层分类器在训练数据的过程中时间消耗较多，而且随着数据量的增大，其在时间损耗方面的缺陷更加明显。相反，以 Softmax 作为网络顶层分类器的深度玻尔兹曼机模型在训练时间上优势显著，同时其分类准确率也与 DBM+SVM 组合模型相差无几。因此，在面对大体量数据的训练任务时，选择哪一种模型更加合适还要根据任务的时间敏感性来决定。

4 结束语

特征量的表达能力与分类器性能是影响乐器分类结果的两个最主要的因素，而深度学习强大的特征提取能力可以为顶层分类器提供概括性更强的音频特征量，但就目前该领域的研究成果来看，将深度学习技术应用于乐器分类问题的研究相对较少。本文以深度玻尔兹曼机模型作为特征提取器提取 5 类乐器纯音乐文件的深度学习特征量，并在神经网络顶层分别设置 SVM 与 Softmax 作为分类器，构建 DBM+SVM 和 DBM+Softmax 组合模型，并在相同的训练数据集与测试数据集上开展实验，结果显示 DBM+SVM 模型的分类准确率最高，但时间消耗也最大，DBM+Softmax 在训练时间上优势明显且准确率与前者相比差距不大。在数据规模较大且时间敏感性较强的分类任务中，DBM+Softmax 组合模型的应用效果应该更好。文章另外还单独采用传统分类模型 SVM 作为浅层模型的代表与以上两种深度学习组合模型进行分类性能对比研究，结果显示 SVM 分类器对 5 类乐器的识别率明显低于基于深度玻尔兹曼机的分类方法，从而展现了深度学习方法强大的学习与建模能力以及其在乐器分类领域的应用前景。

深度学习在乐器分类领域的应用尚属起步阶段，许多问题有待进一步研究，如样本数据的离群点检测及如何有效提升音频分类模型的稀疏表达等都是未来的研究热点。Zhigang Wang 等人[19]提出了一种增量 MI 离群点检测算法 Inc I-MLOF，可以实现批处理模式下的多实例异常值检测。文献[20]对非监督学习方法 SRC 进行了改进，提出了一种基于层次稀疏表示的分类方法，它将单层稀疏表示增强为深度字典形式的多层次表达，通过在多特征数据集上的训练，验证了其性能优势。上述文献提出的方法也是日后乐器分类研究可以考虑借鉴的新思路。

参考文献：

[1] Marques J, Moreno P J. A study of musical instrument classification using gaussian mixture models and support vector machines [R]. [S. l] : Compaq Corporation Cambridge Research Laboratory, 1999.

[2] Liu J, Xie L. SVM-based automatic classification of musical instruments [C]// Proc of International Conference on Intelligent Computation Technology and Automation. Washington DC: IEEE Computer Society, 2010: 669-673.

[3] Zlatintsi A, Maragos P. Am-fm modulation features for music instrument signal analysis and recognition [C]// Proc of Signal Processing Conference. 2014: 2035-2039.

[4] Zhao ZengShun, Feng Xiang, Wei Fang, et al Learning representative features for robot topological localization [J] International Journal of Advanced Robotic Systems, 2013, 10 (4) .

[5] 王芳. 基于深度学习的音乐流派及中国传统乐器识别分类研究 [D]. 南京: 南京理工大学, 2017. (Wang Fang. A study of the classification of music genre and Chinese traditional instruments based on deep learning. [D]. NanKing: Nanjing University of Science and Technology, 2017)

[6] Hinton G E, Srivastava N, Krizhevsky A, *et al.* Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3 (4): págs. 212-223.

[7] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. [J]. Science, 2006, 313 (5786): 504.

[8] Bengio Y. Learning deep architectures for AI [J]. Foundations and Trends in Machine Learning, 2009, 2 (1): 1-127.

[9] Hastad J, Goldmann M. On the power of small-depth threshold circuits [J]. Computational Complexity, 1991, 1 (2): 113-129.

[10] 郭丽丽，丁世飞. 深度学习研究进展 [J]. 计算机科学, 2015, 42 (3): 29. (Guo LiLi, Ding ShiFei. Research Progress on Deep Learning. [J] Computer Science, 2015, 42 (3): 29.)

[11] Bengio Y, Delalleau O. On the expressive power of deep architectures [C]// Proc of International Conference on Discovery Science. [S. l.] : Springer-Verlag, 2011.

[12] Salakhutdinov R, Hinton G. Deep Boltzmann machines [J]. Journal of Machine Learning Research, 2009, 5 (2): 1967-2006.

[13] 蒋文，齐林. 一种基于深度玻尔兹曼机的半监督典型相关分析算法 [J]. 河南科技大学学报: 自然科学版, 2016, 37 (2): 47-51. (Jiang Wen, Qi Lin. A semi supervised canonical correlation analysis algorithm based on deep Boltzmann machine. [J] Journal of Henan University of Science and Technology: Natural Science, 2016, 37 (2): 47-51.)

[14] 张娟，杨建功，汪西莉. 条件深度玻尔兹曼机人脸图像分割模型 [J]. 小型微型计算机系统, 2017, 38 (5): 1130-1133. (Zang Juan, Yang Jian Gong, Wang XiLi. Conditional Deep Boltzmann Machine Face Image Segmentation Model. [J]. Journal of Chinese Computer Systems, 2017, 38 (5): 1130-1133.)

[15] 肖云鹏，李松阳，刘宴兵. 一种基于社交影响力和平均场理论的信息传播动力学模型 [J]. 物理学报, 2017, 66 (3): 227-239. (Xiao Yun Peng, Li SongYang, Liu YanBing. An information diffusion dynamic model based on social influence and mean field theory [J]. Acta Phys Sin, 2017, 66 (3): 227-239.)

[16] 陈晔，孙建华，高小杰，等. 一种基于平均场的无线自组织网络时钟同步方法 [J] 计算机学报, 2016, 39 (5): 893-904. (Chen Wu, Sun JianHua,

Gao XiaoJie, *et al.* A clock synchronization method for Ad hoc networks based on mean field. [J] Chinese Journal of Cumputers, 2016, 39 (5): 893-904.)

[17] 李泉林, 樊瑞娜, 许良. 大型自行车共享系统的平均场极限理论与排队模型研究 [J]. 运筹与管理, 2017, 26 (6): 107-116. (Li Quan lin, Fan RuiNa, Xu Liang. Research on mean field limit theory and queuing model for large-scale bike-sharing systems [J]. Operations Research and Management Science, 2017, 26 (6): 107-116)

[18] Swanston D J, Bishop J M, Mitchell R J. Simple adaptive momentum: new algorithm for training multilayer perceptrons [J]. Electronics Letters, 1994, 30 (18): 1498-1500.

[19] Wang Zhigang, Zhao Zengshun, Weng Shifeng. et al, Incremental multiple instance outlier detection [J]. Neural Computing & Applications, 2015, 26 (4) 957-968.

[20] Wang Zhengxia, Teng Shenghua, Liu Guodong, et al, Hierarchical sparse representation with deep dictionary for multi-modal classification. [J]. Neurocomputing, 2017, 253 (C): 65-69.